KMD – Scientific Team Project SoSe24

Stream and Feature Acquisition Visualization

- Aditya Ganesh Khedekar
- Mallika Manam
- Shreeya Channappa Yogesh

SUPERVISOR: Mr. Maik Büttner

Agenda

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
INF FAKULTÄT FÜR INFORMATIK

- Introduction
- Motivation
- Data Drift
- Visualization of Missing Data
- Velocity
- Missing Data Analysis
- Concept Drift
- Learning Strategies

Introduction

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG INF FAKULTÄT FÜR INFORMATIK

➔ **Stream Data Visualization:** Visualization of continuously generated and real-time processed data.

➔ **Goals:**

- Research stream visualisation methods.
- Implement at least one method for concept drift detection.
- Implement visualisations for concept and feature drift.
- Visualise the velocity of the stream.
- Explore ways to visualize missing values.
- Implement visualisation for comparing performance metrics of different learning strategies over the course of the stream.

➔ **Achieved by :**

- Integrating the above implementations into a python package in an object-oriented way .

Introduction

➔ **Datasets:**

- **cfpdss.csv** :
  - Synthetically generated dataset
  - 10 features (5 numerical and 5 categorical)
  - Categorical Target with binary variable.

- **cfpdss_m0.5.csv** : cfpdss dataset with missing values.

- **experiment.csv** :
  - Contains seven strategies/models
  - Dataset is divided into batch. Each batch has 50 instances.
  - For each strategy, the kappa score is given on for each batch.

Motivation

Why visualize?

➔ Monitor system performance and improve it accordingly.

➔ Faster data exploration and decision making.

➔ Improved operational efficiency.

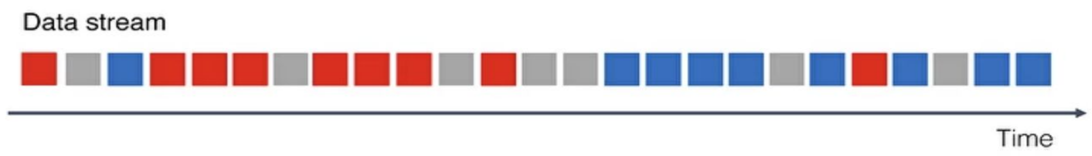


Fig 1: Depiction of a data stream over time with different concepts depicted by different colours [1]

Data Drift

- Citation [6] & [10]

➔ Virtual/Data Drift: Changes in the input distribution p(X) and change in distribution of the label p(y).

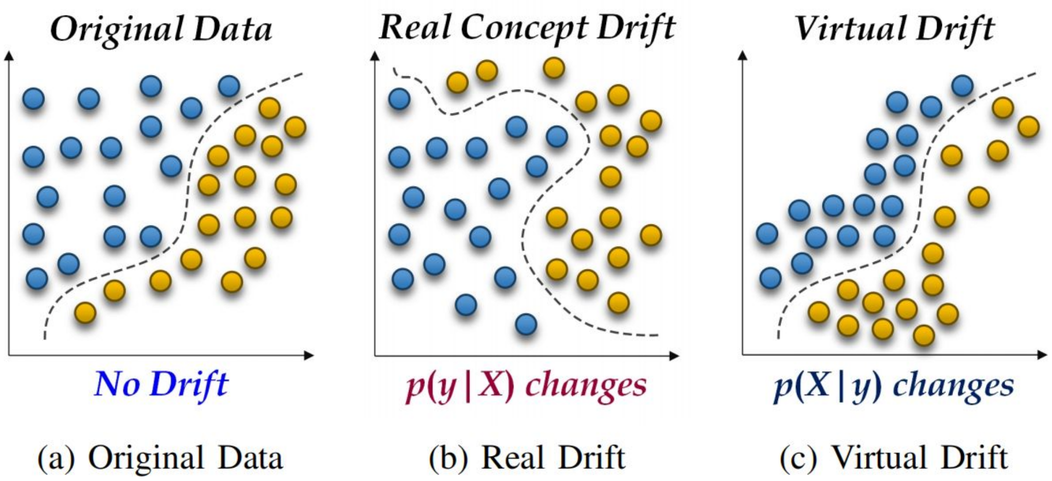


Fig 2: Figure depicting real concept drift and virtual drift [10]

Data Drift

Types of Drifts:

- Citation [6] & [10]

1. Incremental / Linear Drift
    - Incremental consisting of many intermediate concepts in between
    - Sequence of data distributions appear during the transition
    - Eg, a sensor slowly wears off and becomes less accurate
2. Gradual Drift
    - Gradual concept drift results from a slow transition from one data distribution to the next.
    - Eg, relevant news topics change from dwelling to holiday homes, while the user does not switch abruptly, but rather keeps going back to the previous interest for some time
3. Sudden / Abrupt Drift
    - An abrupt concept drift results from a sudden change in the data distribution
    - Eg, replacement of a sensor with another sensor that has a different calibration in a chemical plant

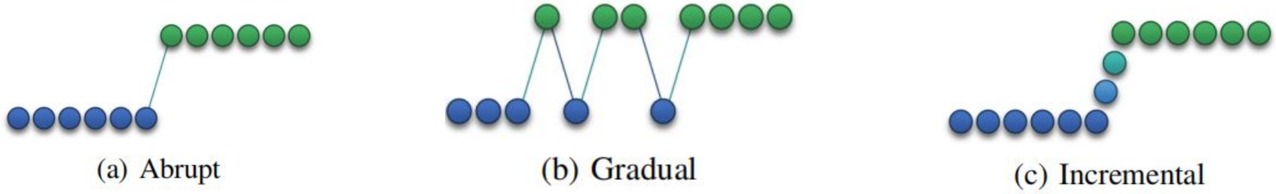


Fig 2: Figures to show abrupt, gradual and incremental drifts [6]

Data Drift

- Windowing Technique: Sliding Window (Dequeue)

- Drift Detection Technique:
  - Kolmogorov - Smirnov (KS) Test - numerical features
  - Population Stability Index (PSI) Test - categorical features

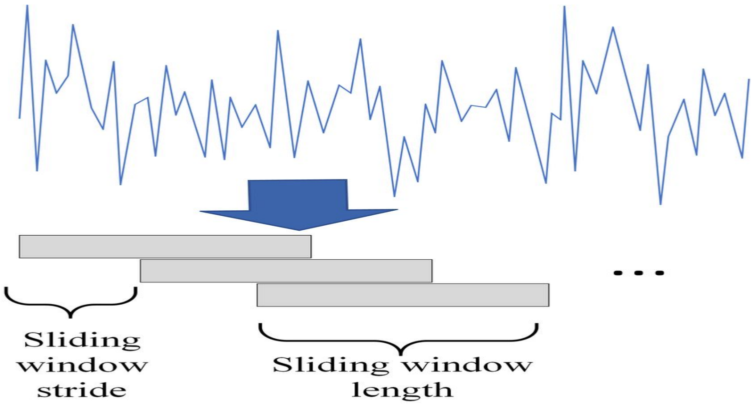


Fig 3: Figure depicting sliding window technique [3]

Data Drift

- Conditions to detect different types of drifts - the conditions are checked only if the p-value is below the significance level for KS test and psi-value is greater than the set threshold for PSI test.
  - Sudden Drift: $abs(mean_{diff}) > std(window)$

    $$mean_{diff} = mean(second\ half\ of\ window) - mean(first\ half\ of\ window)$$

  - Linear Drift: $mean_{diff} > 0$
  - Gradual Drift: Change in windowing technique, introduction of a gap in between the two halves of the windows [4].

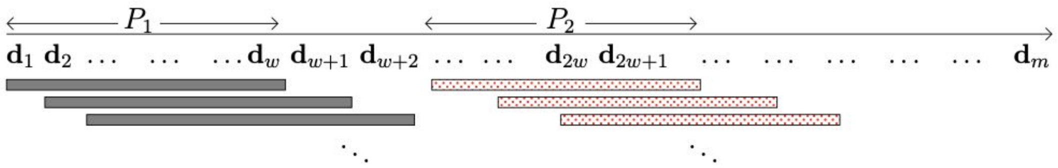


Fig 4: Sliding window with gap [4]

Data Drift

Feature Drift Visualization

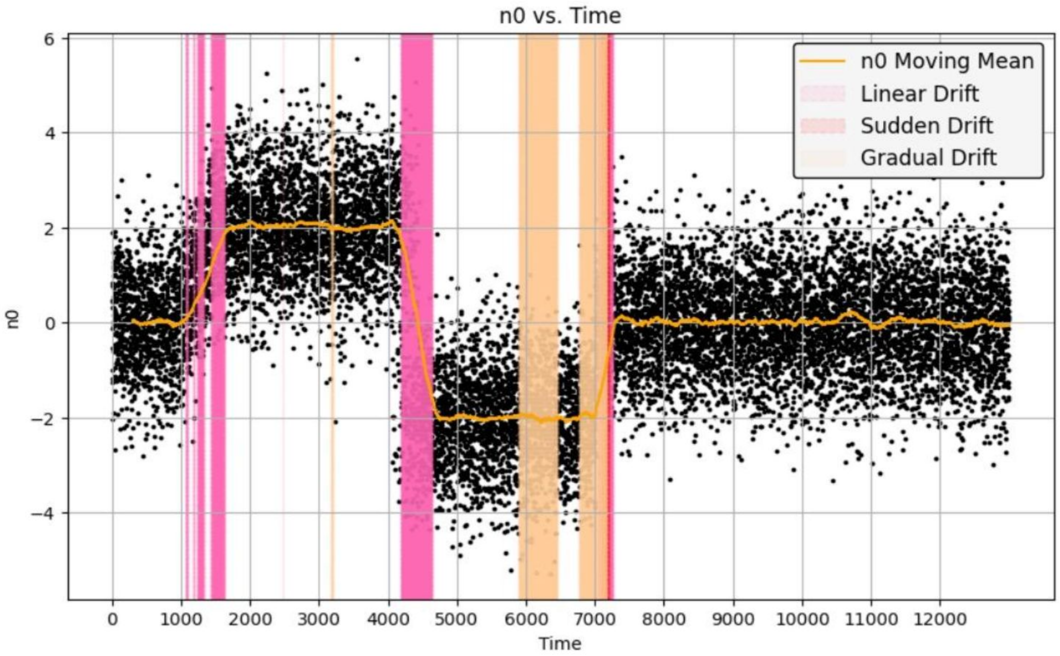


Fig 5: Graph depicting sudden, linear and gradual drift with window_size = 300 and gap_size = 100 for the numerical feature ‘n0

Data Drift

Feature Drift Visualization

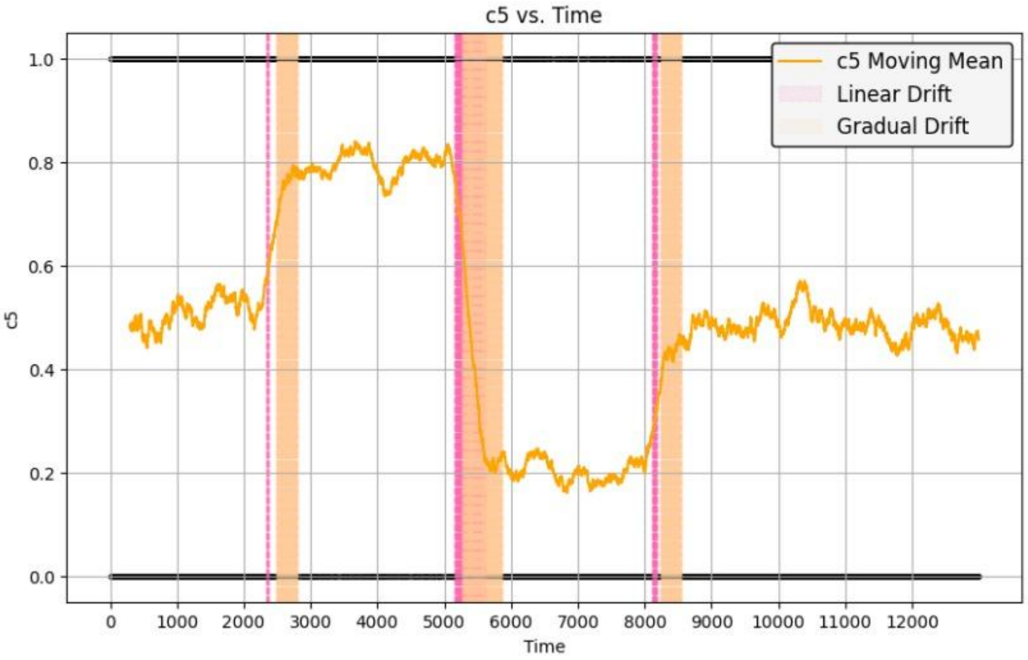


Fig 6: Graph depicting linear and gradual drift for the categorical feature ‘c5’ with window_size 300 and gap_size 100

Missing Data

Visualization of Missing Data

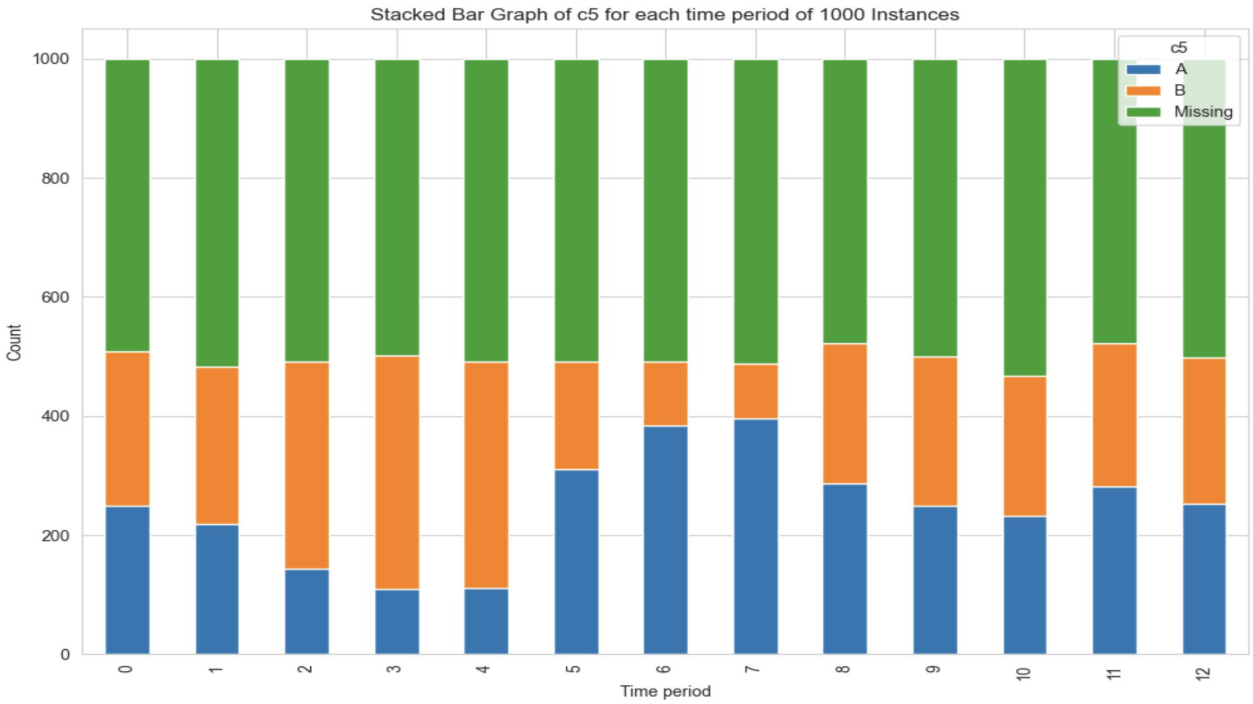


Fig 7: Stacked bar chart depicting the number of missing data in the categorical feature ‘c5’

Missing Data

Visualization of Missing Data

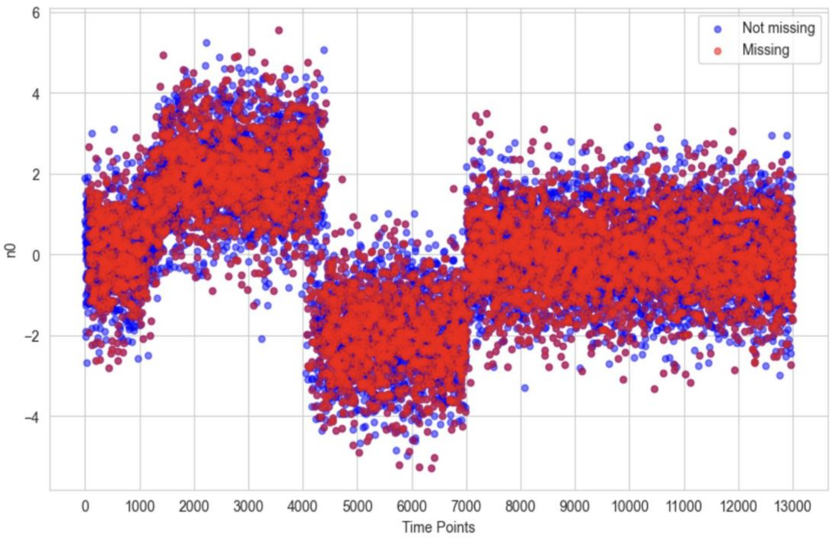


Fig 8: Scatter plot depicting missing and non-missing values for the numerical feature ‘n0’

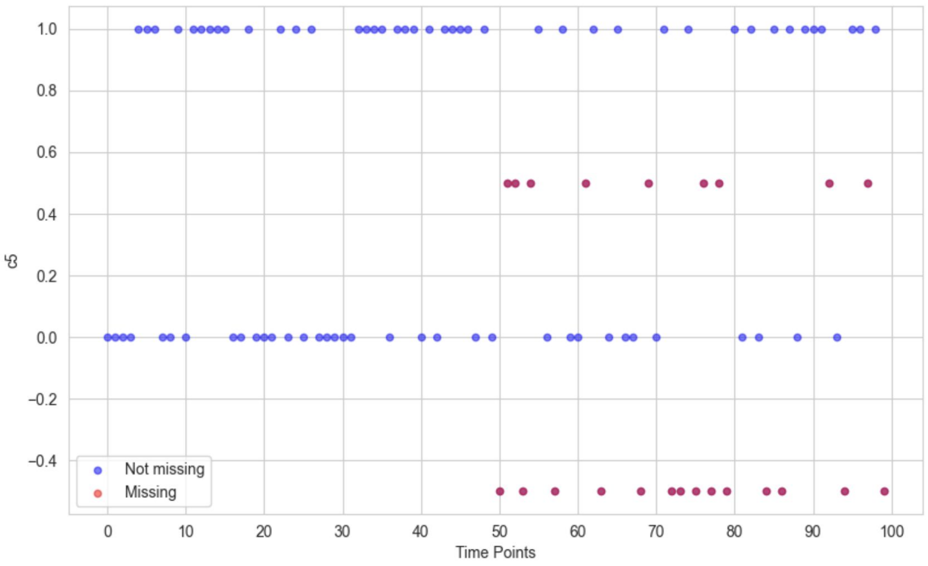


Fig 9: Scatter plot depicting missing and non-missing data in the categorical feature ‘c5’

Missing Data

Visualization of Missing Data

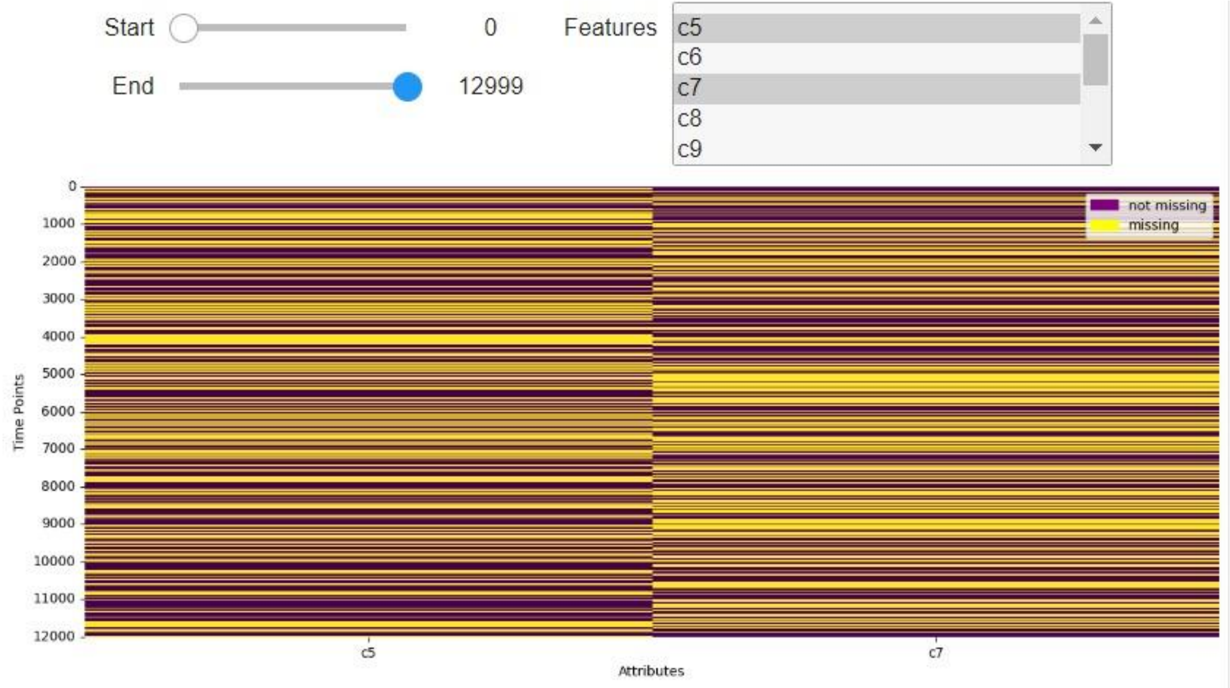


Fig 10: Heatmap with sliders depicting missing data

Velocity

Data Velocity: It is the rate at which data is generated and processed within a system[12].

Visualization of data velocity

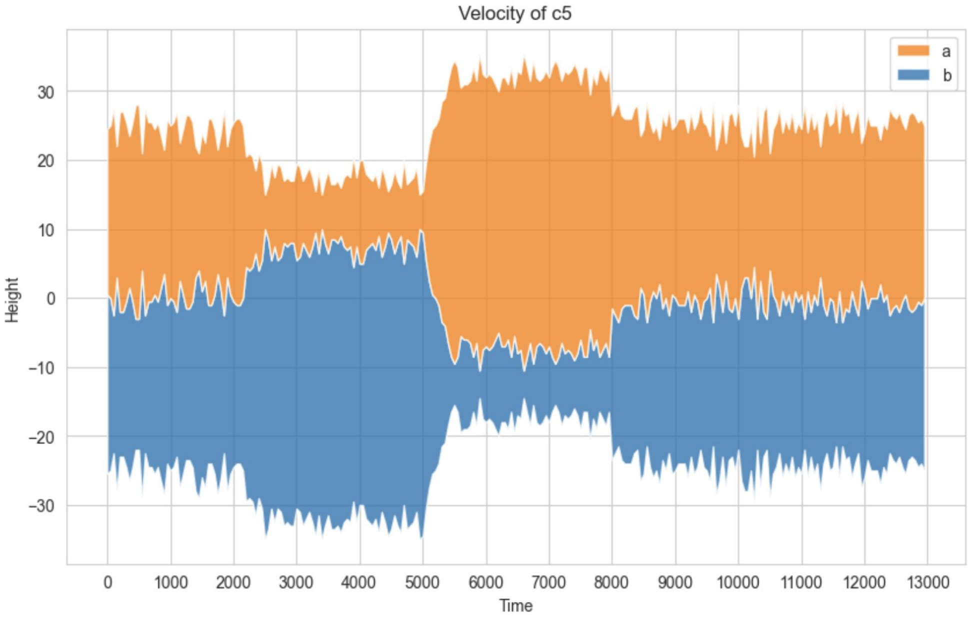
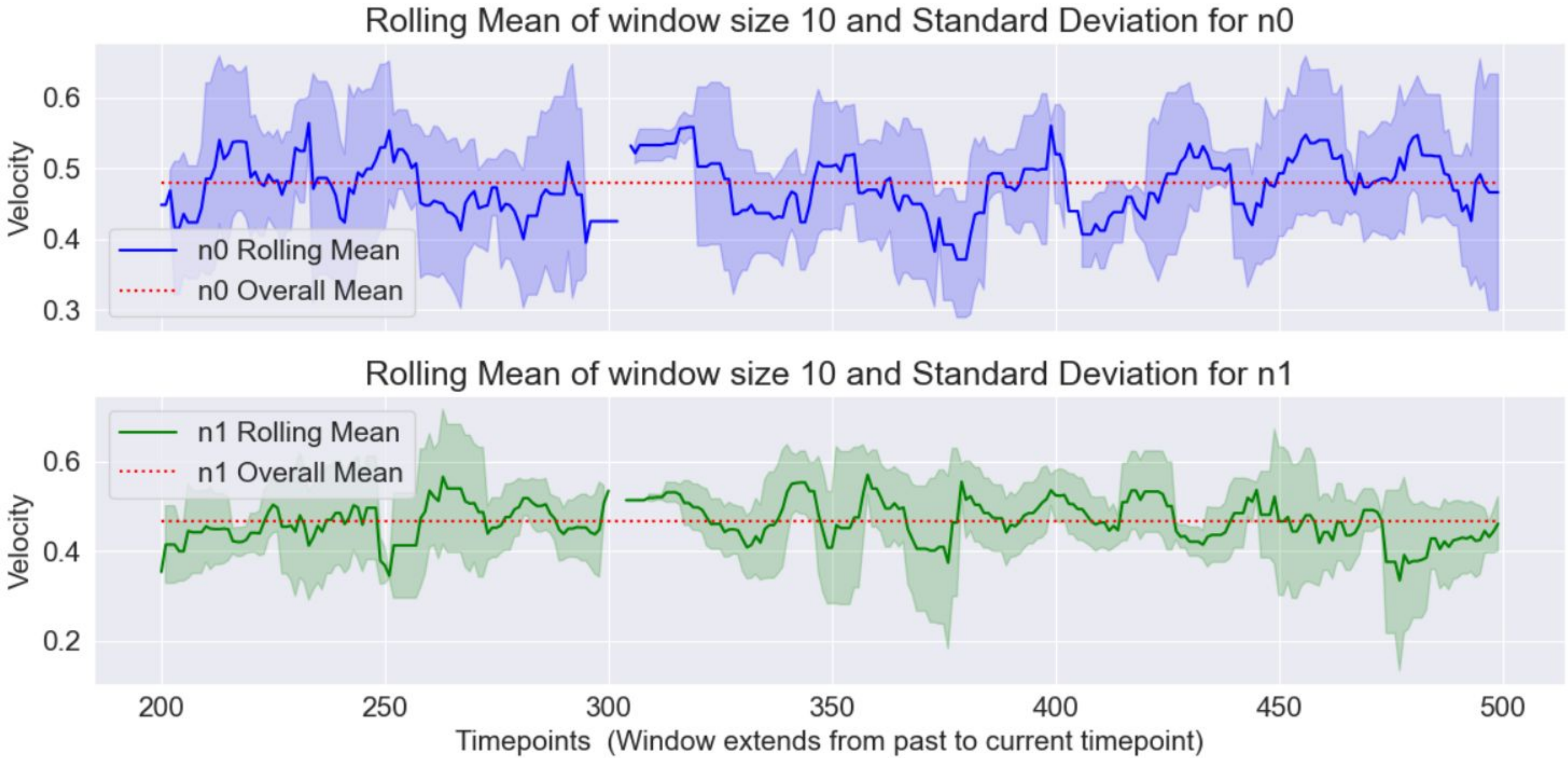


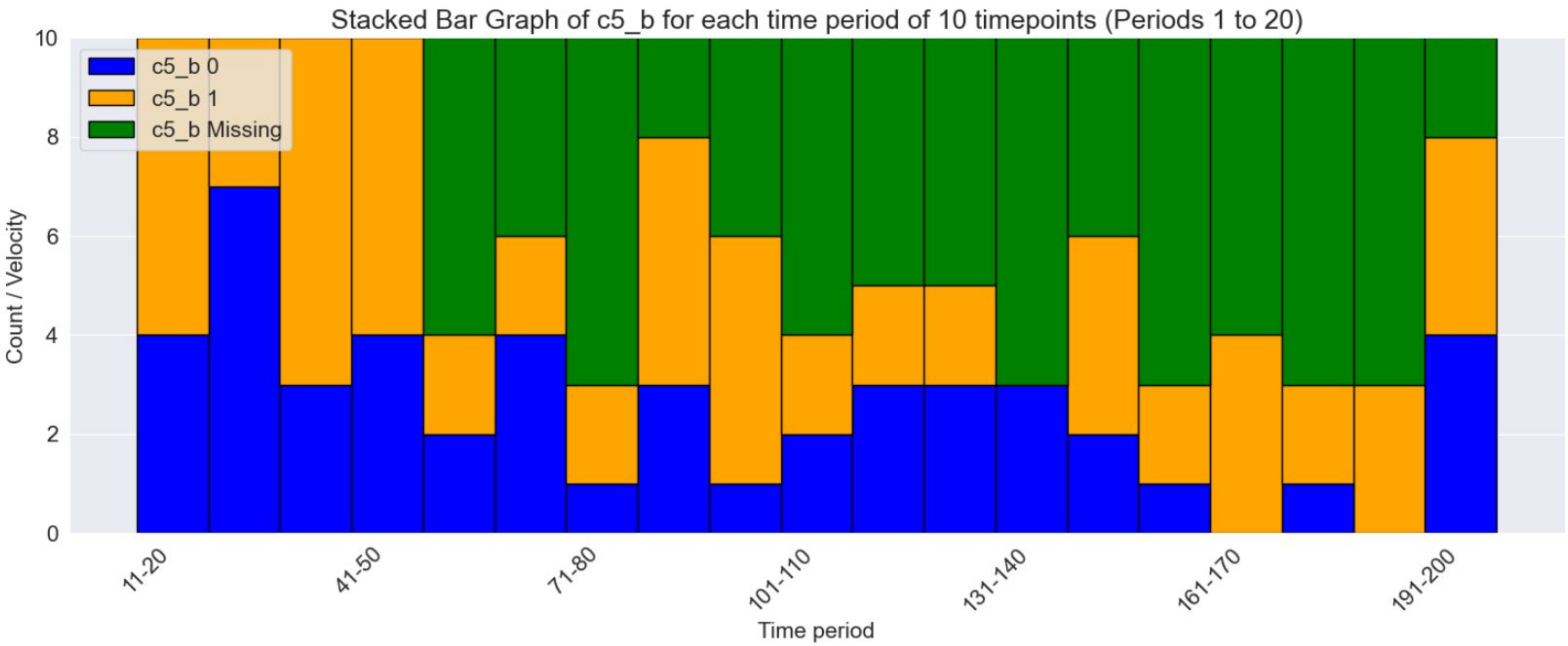
Fig 11: Stream Graph for categorical feature ‘c5’ with bin size = 50

Velocity



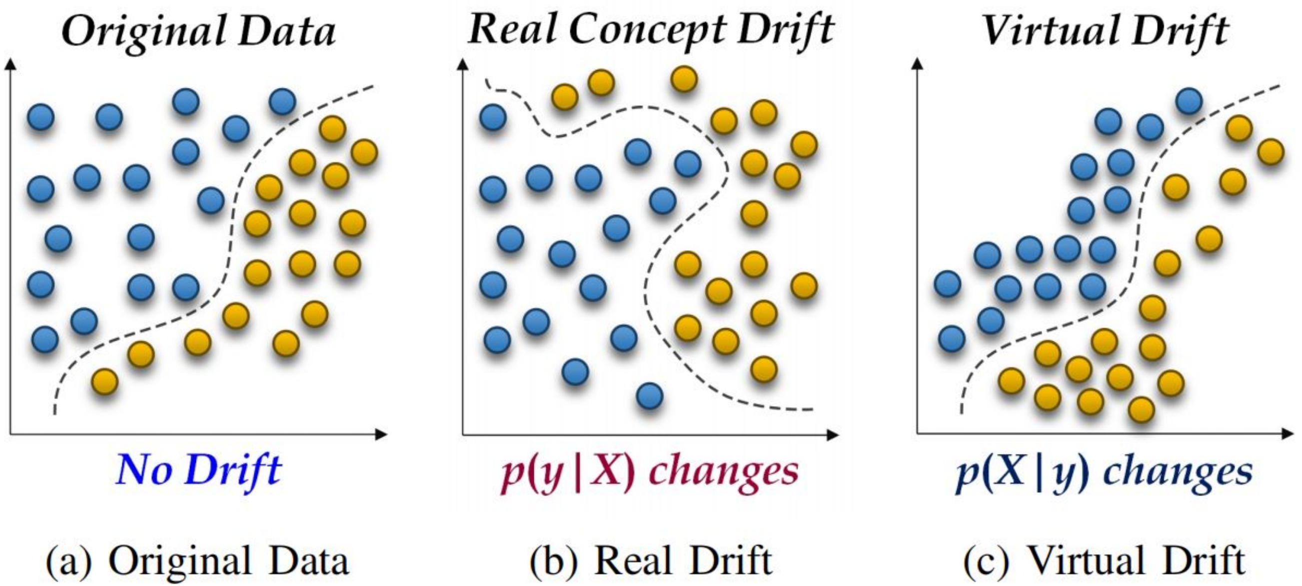
Rolling Mean of window size 10 and Standard Deviation for n0
Velocity
0.6
0.5
0.4
0.3
n0 Rolling Mean
n0 Overall Mean
Rolling Mean of window size 10 and Standard Deviation for n1
Velocity
0.6
0.4
0.2
n1 Rolling Mean
n1 Overall Mean
200
250
300
350
400
450
500
Timepoints (Window extends from past to current timepoint)

Velocity


Stacked Bar Graph of c5_b for each time period of 10 timepoints (Periods 1 to 20)
c5_b 0
c5_b 1
c5_b Missing
Count / Velocity
Time period
11-20
41-50
71-80
101-110
131-140
161-170
191-200
0
2
4
6
8
10

Real Concept Drift

- Citation [6] & [10]



A real concept drift refers to the changes in p(y|X) which affects the decision boundaries or the target concept

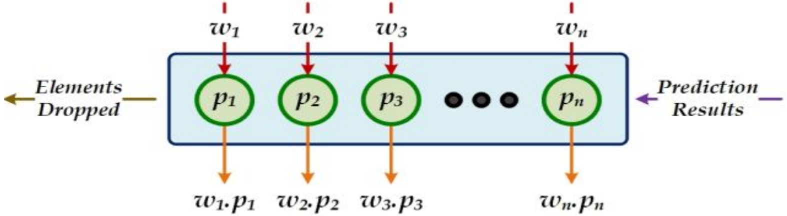
Initially user was interested in news articles related to dwelling houses, but now interested in holiday homes.

McDiarmid Drift Detection Method (MDDM)

- Citation [6] & [10]

- MDDM applies McDiarmid's inequality to detect concept drifts

- Sliding Window Approach
- 1 for correct prediction, 0 otherwise



- Weighting scheme for element in window : $w_i < w_{i+1}$
  - Arithmetic : $w_i = 1 + (i + d)$ ...where d ≥ 0, is difference between two consecutive weights
  - Geometric : $w_i = r^{(i-1)}$ ...where r ≥ 1, is ratio between two consecutive weights
  - Euler: $w_i = r^{(i-1)}$ with $r = e^{\lambda}$ ... where $\lambda \geq 0$

- McDiarmid's inequality is calculated as follows: $\varepsilon_w = \sqrt{\frac{\sum_{i=1}^{n} v_i^2}{2} \ln \frac{1}{\delta_w}}$
  - where, n is number of entries in window and $v_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$
  - δ_w is the confidence level

McDiarmid Drift Detection Method

- Two variables tracked
  - Weighted average of the elements of the sliding window, μ_w^t
  - Maximum weighted mean observed so far, μ_w^m

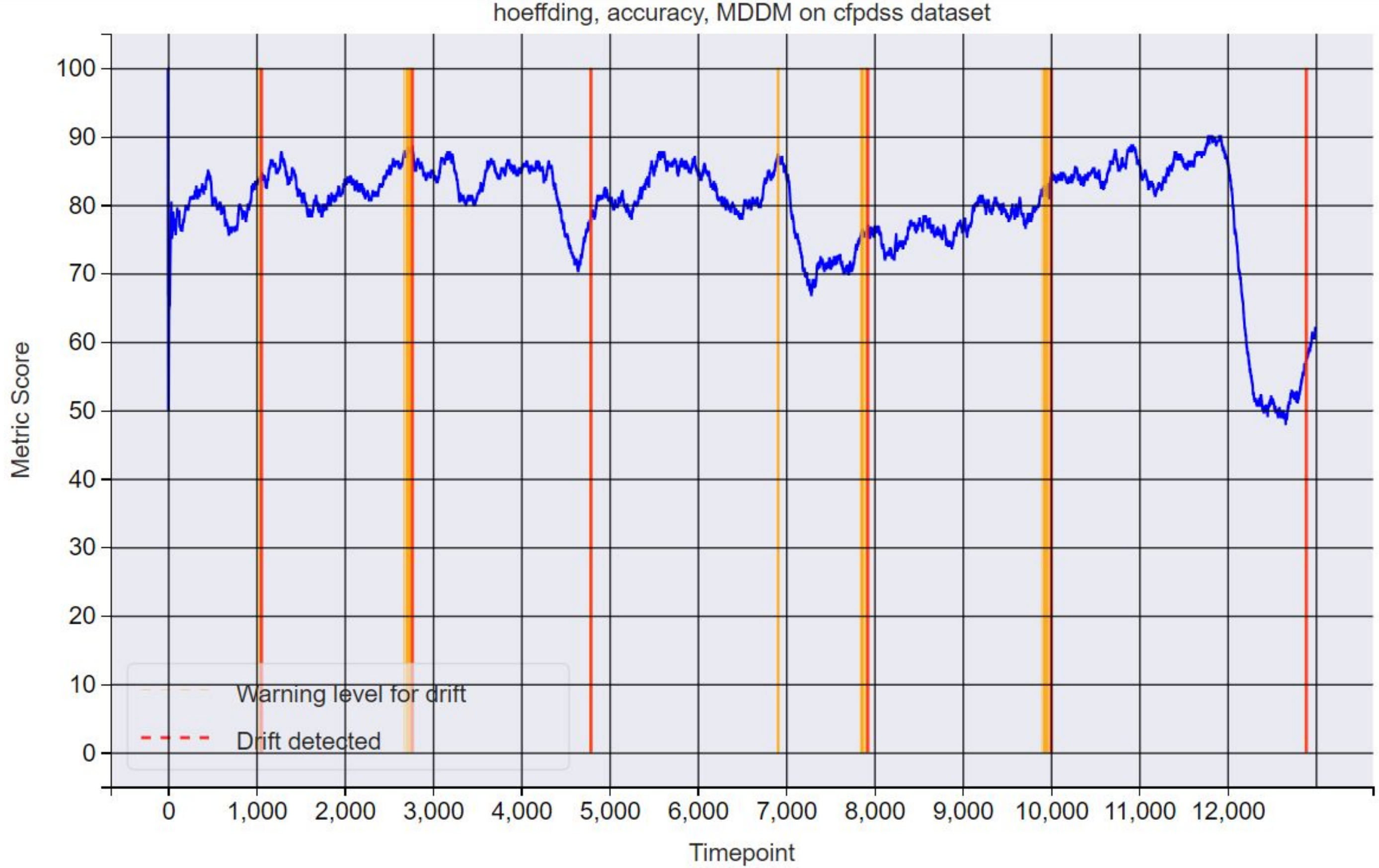
- Ideally, Accuracy (or metric) should increase or stay constant over time as number of instances increases

- Possibility of facing a concept drift increases if μ_w^m does not change and μ_w^t decreases over time.

- Drift detected when : $\mu_w^m - \mu_w^t \geq \varepsilon_d$ ...where ε_d is McDiarmid Inequality

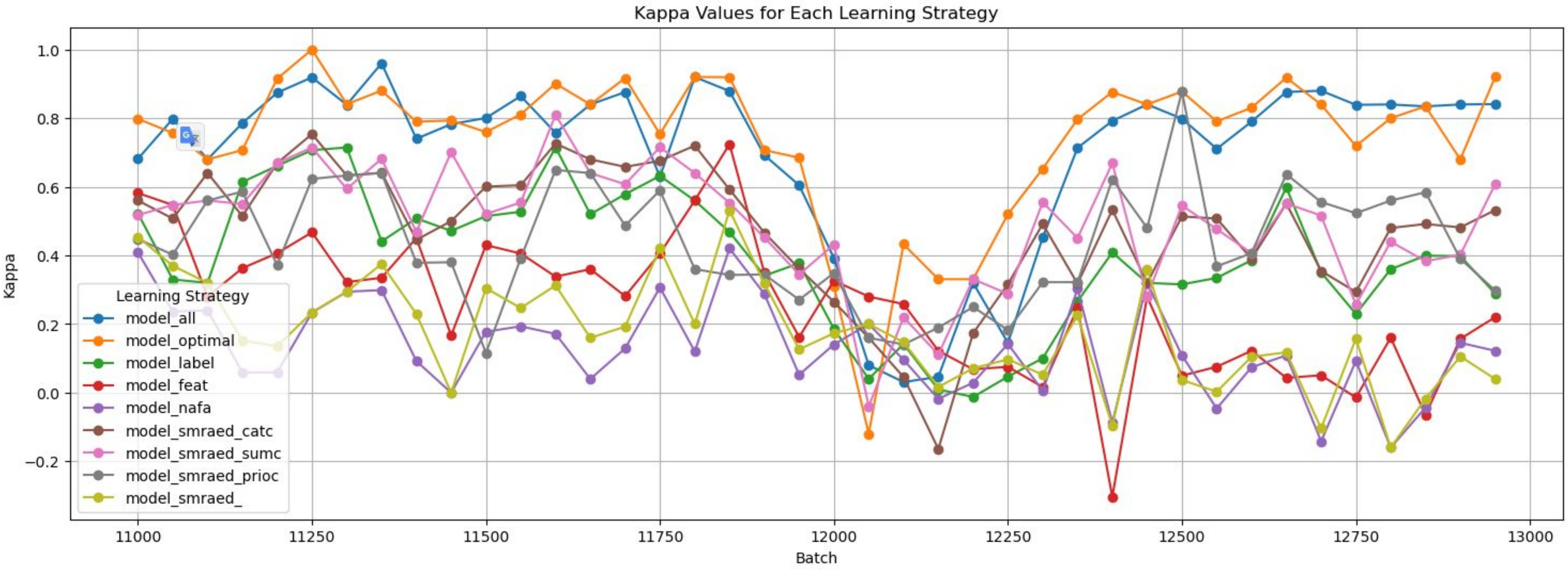
- Optimal values: $\delta_w = 10^{-6}$, d = 0.01, r = 1.01, λ = 0.01.

MDDM - Arithmetic scheme

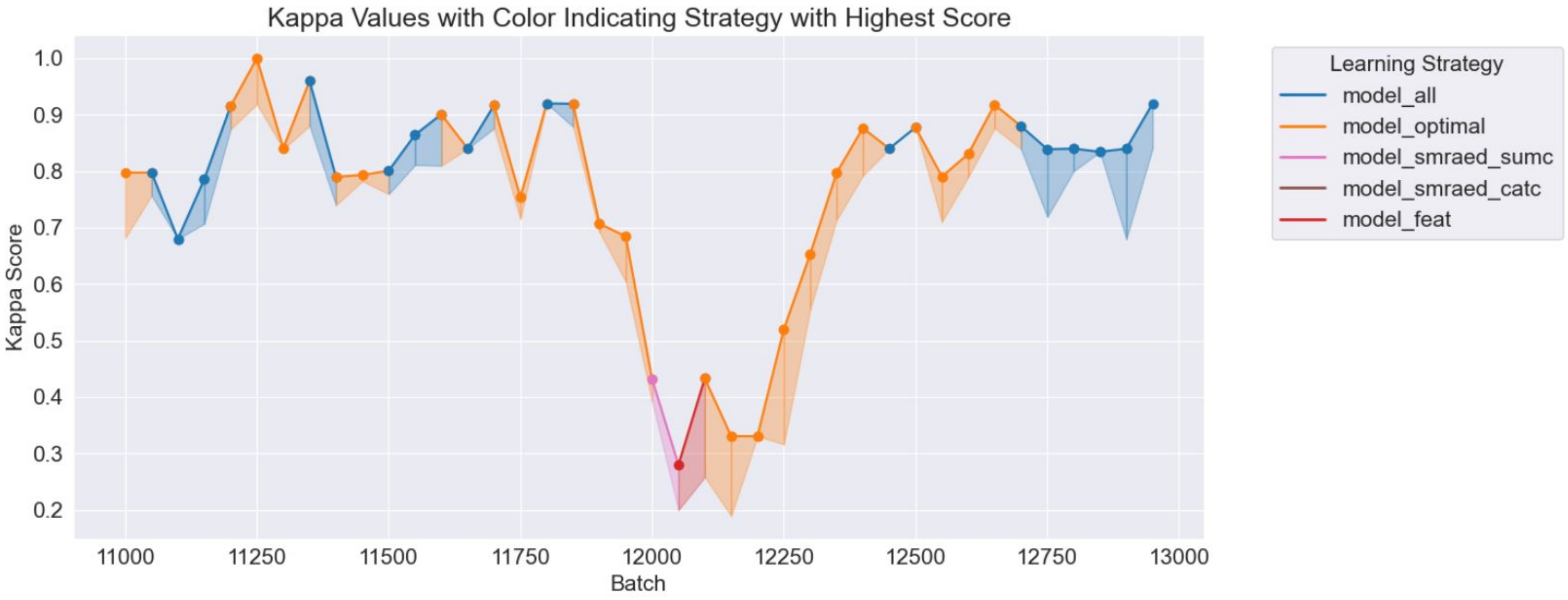


Real Concept Drift

Learning Strategies



Learning Strategies



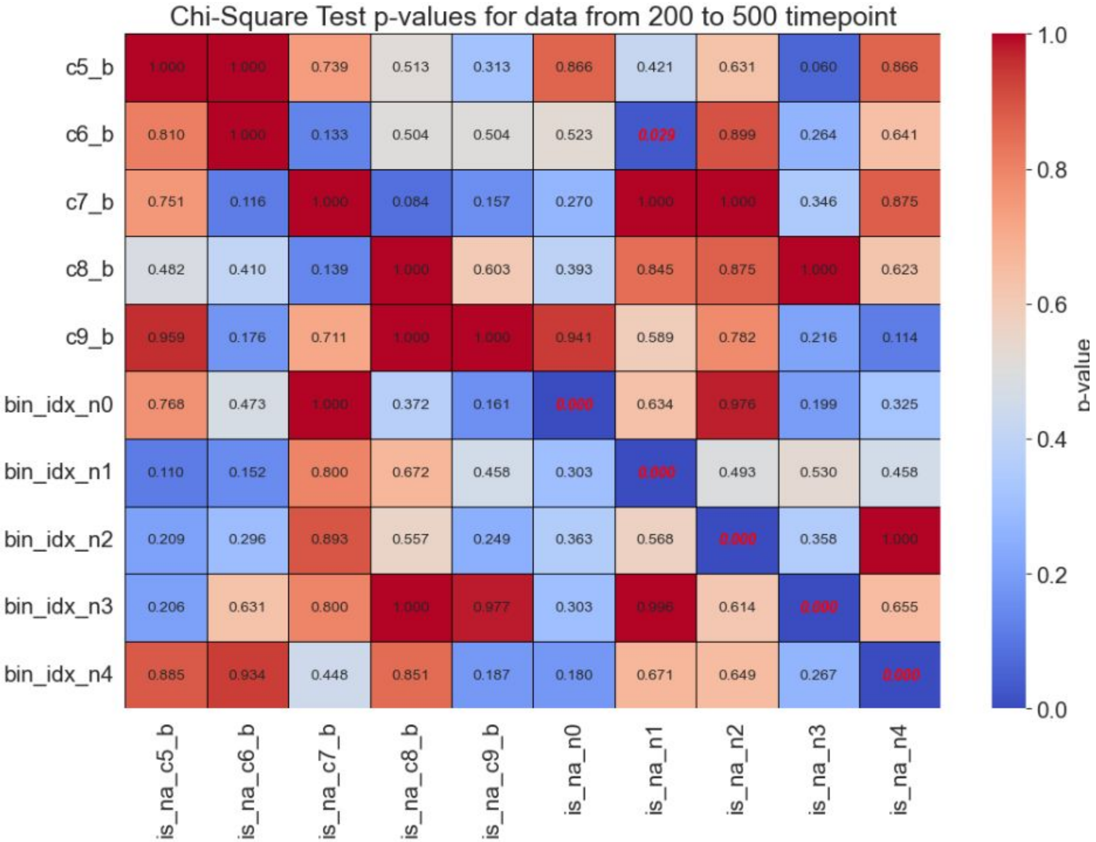
Data Missingness

- Citation [14]

1. MCAR (Missing Completely at Random):
   - Little MCAR Test
   - Null hypothesis: Data is Missing Completely At Random (MCAR).
   - If p-value greater than significance level (0.05), we fail to reject null hypothesis.

2. MAR (Missing at Random):
   - Binned numerical features with help of decision trees
   - Add extra columns in dataset to indicate presence of missing value (*is_na_col*)
   - Used Chi-Square to check for dependency between feature and `*is_na_col*` columns
   - Null Hypothesis: No relationship between given two variables
   - Assumed Significance level as 0.05
   - If p-value is greater than the significance level, we fail to reject null hypothesis indicating data is not MAR

Data Missingness



References

1. https://deepchecks.com/data-drift-vs-concept-drift-what-are-the-main-differences

2. https://towardsdatascience.com/understanding-kolmogorov-smirnov-ks-tests-for-data-drift-on-profiled-data-5c8317796f78

3. Deep Learning for Load Forecasting with Smart Meter Data: Online Adaptive Recurrent Neural Network - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Sliding-window-technique_fig2_346510102 [accessed 16 Jun, 2024]

4. Martjushev, J. & R.P., Jagadeesh Chandra Bose & Aalst, Wil. (2015). Change Point Detection and Dealing with Gradual and Multi-order Dynamics in Process Mining. 161-178. 10.1007/978-3-319-21915-8_11.

5. Krstajić, Miloš & Keim, Daniel. (2013). Visualization of streaming data: Observing change and context in information visualization techniques. Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013. 41-47. 10.1109/BigData.2013.6691713.

6. Pesaranghader, A., Viktor, H.L., & Paquet, E. (2017). McDiarmid Drift Detection Methods for Evolving Data Streams. 2018 International Joint Conference on Neural Networks (IJCNN), 1-9.

7. M. Krstajić and D. A. Keim, "Visualization of streaming data: Observing change and context in information visualization techniques," 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 2013, pp. 41-47, doi:10.1109/BigData.2013.6691713.

References

8. Chakrabarti, Arnab; Kulshrestha, Tanuj; Quix, Christoph, "A Visualization System for High Dimensional Data Streams using Complex Event Processing", Information Visualization of Geospatial Networks, Flows and Movement in conjugation with IEEE VIS2020, MoVIS2020

9. https://medium.com/@ayeshasidhikha188/types-of-missing-values-fba155099ac7

10. João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. ACM Comput. Surv. 46, 4, Article 44 (April 2014), 37 pages. https://doi.org/10.1145/2523813

11. Palmeiro, João and Malveiro, Beatriz and Costa, Rita and Polido, David and Moreira, Ricardo and Bizarro, Pedro. 2022. Data+Shift: Supporting Visual Investigation of Data Distribution Shifts by Data Scientists. EuroVis 2022 - Short Papers, The Eurographics Association. https://doi.org/10.2312/evs.20221097

12.https://www.dremio.com/wiki/data-velocity/#:~:text=Data%20Velocity%20refers%20to%20the,through%20a%20system%20or%20organization.

13. https://www.machinelearningplus.com/deployment/population-stability-index-psi/

14.Roderick J. A. Little. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of the American Statistical Association, 83(404), 1198–1202. https://doi.org/10.2307/2290157

Thank you!

Questions?